# REPORT DOCUMENTATION PAGE

*Form Approved*
**OMB No. 0704-0188**



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| MARCH 2012 | Conference Paper (PREPRINT) | |

**4. TITLE AND SUBTITLE**

INTEGRATED FEATURE NORMALIZATION AND ENHANCEMENT FOR ROBUST SPEAKER RECOGNITION USING ACOUSTIC FACTOR ANALYSIS (PREPRINT)

**5a. CONTRACT NUMBER**
FA8750-09-C-0067

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
35885G

**6. AUTHOR(S)**

Taufiq Hasan, and John H. L. Hansen

**5d. PROJECT NUMBER**
3188

**5e. TASK NUMBER**
BA

**5f. WORK UNIT NUMBER**
AE

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Research Associates for Defense Conversion, Inc.
10002 Hillside Terrace
Marcy NY 13403

Subcontractor: University of Texas at Dallas
800 W Campbell Rd
Richardson, TX 75080

**8. PERFORMING ORGANIZATION REPORT NUMBER**
N/A

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Air Force Research Laboratory/Information Directorate
Rome Research Site/RIGC
525 Brooks Road
Rome NY 13441-4505

**10. SPONSOR/MONITOR'S ACRONYM(S)**
N/A

**11. SPONSORING/MONITORING AGENCY REPORT NUMBER**
AFRL-RI-RS-TP-2012-040

**12. DISTRIBUTION AVAILABILITY STATEMENT**

Approved for public release; distribution unlimited. PA# 88ABW-2012-1819
Cleared Date: 29 March 2012

**13. SUPPLEMENTARY NOTES**

This paper was accepted for publication in the Proceedings of the 2012 Interspeech Conference, Portland, Oregon, 9-13 Sept-2012. This work was funded in whole or in part by Department of the Air Force contract number FA8750-09-C-0067. The U.S. Government has for itself and others acting on its behalf an unlimited, paid-up, nonexclusive, irrevocable worldwide license to use, modify, reproduce, release, perform, display, or disclose the work by or on behalf of the Government. All other rights are reserved by the copyright owner.

**14. ABSTRACT**

State-of-the-art factor analysis based channel compensation methods for speaker recognition are based on the assumption that speaker/utterance dependent Gaussian Mixture Model (GMM) mean super-vectors can be constrained to lie in a lower dimensional subspace, which does not consider the fact that conventional acoustic features may also be constrained in a similar way in the feature space. In this study, motivated by the low-rank covariance structure of cepstral features, we propose a factor analysis model in the acoustic feature space instead of the super-vector domain and derive a mixture of dependent feature transformation. We demonstrate that, the proposed Acoustic Factor Analysis (AFA) transformation performs feature dimensionality reduction, de-correlation, variance normalization and enhancement at the same time. The transform applies a square-root Wiener gain on the acoustic feature eigenvector directions, and is similar to the signal sub-space based speech enhancement schemes. We also propose several methods of adaptively selecting the AFA parameter for each mixture. The proposed feature transformation is applied using a probabilistic mixture alignment, and is integrated with a conventional i-Vector system. Experimental results on the telephone trials of the NIST SRE 2010 demonstrate the effectiveness of the proposed scheme.

**15. SUBJECT TERMS**

Channel Normalization, Factor Analysis, Tactical SIGINT Technology, Gaussian Modeling

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | JOHN G. PARKER |
| U | U | U | UU | 5 | 19b. TELEPHONE NUMBER (*Include area code*) N/A |

# Integrated Feature Normalization and Enhancement for robust Speaker Recognition using Acoustic Factor Analysis

*Taufiq Hasan and John H. L. Hansen**

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.

{Taufiq.Hasan, John.Hansen}@utdallas.edu

## Abstract

State-of-the-art factor analysis based channel compensation methods for speaker recognition are based on the assumption that speaker/utterance dependent Gaussian Mixture Model (GMM) mean super-vectors can be constrained to lie in a lower dimensional subspace, which does not consider the fact that conventional acoustic features may also be constrained in a similar way in the feature space. In this study, motivated by the low-rank covariance structure of cepstral features, we propose a factor analysis model in the acoustic feature space instead of the super-vector domain and derive a mixture dependent feature transformation. We demonstrate that, the proposed Acoustic Factor Analysis (AFA) transformation performs feature dimensionality reduction, de-correlation, variance normalization and enhancement at the same time. The transform applies a square-root Wiener gain on the acoustic feature eigenvector directions, and is similar to the signal sub-space based speech enhancement schemes. We also propose several methods of adaptively selecting the AFA parameter for each mixture. The proposed feature transform is applied using a probabilistic mixture alignment, and is integrated with a conventional i-Vector system. Experimental results on the telephone trials of the NIST SRE 2010 demonstrate the effectiveness of the proposed scheme.

## 1. Introduction

Factor analysis based channel compensation methods for speaker recognition are based on the assumption that, speaker/utterance dependent adapted GMM [1] mean super-vectors can be constrained to lie in a lower dimensional subspace [2–4]. Lower dimensional speaker and channel dependent subspaces assumption inspired various channel compensation schemes such as, Eigenvoice [2], Eigenchannel [3] and Joint Factor Analysis (JFA) [3]. With the introduction of i-Vectors, which are the *latent factors* of the so called "total variability" space [4], research trend shifted towards directly applying compensation techniques on these lower dimensional utterance level features, enabling the development of fully Bayesian techniques [5, 6].

While super-vector domain factor analysis techniques and its derivatives are effective, they do not consider the fact that acoustic feature vectors can also be constrained in a lower dimensional subspace of the feature space. This is clear, since a low-rank modeling of the super-covariance matrix obtained from different utterance GMMs is not equivalent to a low-rank assumption of the acoustic feature covariance matrix. Lower dimensional representation of speech short-time spectrum is a well established phenomenon which motivated a family of speech enhancement methods known as the *signal subspace* approach [7]. This phenomenon is also found to be valid for popular acoustic features, such as Mel-frequency Cepstral Coefficients (MFCC) [8], even though these features are processed by Discrete Cosine Transform (DCT) for de-correlation. To illustrate that acoustic feature covariance matrices have close to zero eigenvalues, and can be assumed low-rank, we train a 1024 mixture full-covariance GMM Universal Background Model (UBM) using 60 dimensional MFCC features on a large development data set[1]. For a typical mixture of this UBM, the covariance matrix is plotted as an intensity image in Fig. 1(a), revealing that the non-diagonal components are indeed significant. Fig. 1(b) shows the sorted eigenvalues of three different mixtures showing how the energy is compacted in the first few coefficients, while the later ones are close to zero. Also, it is known that the first few eigen-directions of the feature covariance matrix are relatively more speaker dependent [8], further justifying the low-rank assumption of features for a speaker recognition task.

Inspired by the abovementioned observations, in this study, we propose an *acoustic factor analysis* [9] scheme for speaker recognition and develop a mixture-dependent feature dimensionality reduction transform. The proposed transformation performs dimensionality reduction, de-correlation, feature variance normalization, and enhancement, at once. Also, instead of hard feature alignment to a specific mixture, applying the transformation, and then retraining the UBM, we use a probabilistic frame alignment and transform the UBM parameters within the system. Integrating the proposed method within a standard i-Vector system provides significant gain in system performance.

## 2. Acoustic Factor Analysis

In this section, we describe the proposed factor analysis model of acoustic features, discuss its formulation, mixture-wise application for dimensionality reduction, advantages and properties.

### 2.1. Formulation

Let $\mathcal{X} = \{\mathbf{x}_n | n = 1 \cdots N\}$ be the collection of all acoustic feature vectors from the development set. Using a factor analysis model, a $d \times 1$ feature vector $\mathbf{x} \in \mathcal{X}$ can be represented by,

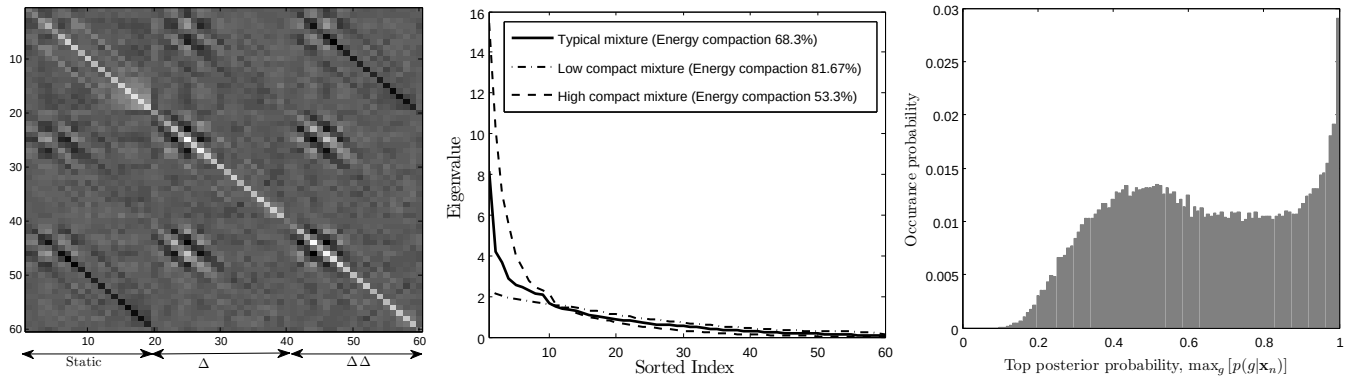
$$\mathbf{x} = \mathbf{W}\mathbf{y} + \mu + \mathbf{c}. \quad (1)$$

Here, $\mathbf{W}$ is a $d \times q$ low rank factor loading matrix that represents $q < d$ bases spanning the subspace with important variability in the feature space, and $\mu$ is the $d \times 1$ mean vector of $\mathbf{x}$. We denote the latent variable vector $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as *acoustic factors*, which is of dimension $q \times 1$. We assume the remaining noise component $\mathbf{c} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ to be isotropic and the model is thus equivalent to Probabilistic Principal Component Analysis (PPCA) [10].

The advantage of this model is that the *acoustic factors* $\mathbf{y}$, explains the correlation between the feature coefficients $\mathbf{x}$, which we believe are more speaker dependent [8], while the noise component $\mathbf{c}$ incorporates the residual variance of the data. It should be em-

*This project was funded by AFRL under contract FA8750-12-1-0188 (Approved for public release, distribution unlimited: 88ABW-2012-1810), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

[1] Details on feature extraction and UBM data are given in Sec. 4.1.




(a) (b) (c)

Figure 1: (a) An intensity image plot showing a typical covariance matrix of a full-covariance UBM (darker indicate lower values) trained on 60-dimensional MFCC features (20 static+Δ+ΔΔ). (b) Sorted eigenvalues of three covariance matrices from the UBM showing different energy compaction in various mixtures. Energy compaction is the percentage of top eigenvectors that account for 95% of total energy. A typical, low and high compact mixture eigenvalues are shown. (c) Distribution of top UBM mixture posterior probability values for development features revealing that frame alignment is not always definitive.

phasized that even though we denote the term $\mathbf{c}$ as "noise", when $\mathbf{x}$ represent cepstral features $\mathbf{c}$ actually represent convolutional channel distortion. Using a mixture of these models [10], we have,

$$p(\mathbf{x}) = \sum_g w_g p(\mathbf{x}|g) \text{ where} \quad (2)$$

$$p(\mathbf{x}|g) = \mathcal{N}(\mu_g, \sigma_g^2\mathbf{I} + \mathbf{W}_g\mathbf{W}_g^T). \quad (3)$$

Here, $\mu_g$, $w_g$, $\mathbf{W}_g$ and $\sigma_g^2$ denote the mean vector, mixture weight, factor loadings, and noise variance for the $g$-th AFA mixture.

### 2.2. Mixture dependent transformation

One advantage of using a mixture of PPCA models for *acoustic factor analysis* is that its parameters can be conveniently extracted from a full-covariance GMM-UBM trained using the Expectation-Maximization (EM) algorithm [10]. The procedure is given below:

#### *2.2.1. Universal Background Model*

The UBM model $\Lambda_0$, trained on the dataset $\mathcal{X}$ is given by,

$$p(\mathbf{x}|\Lambda_0) = \sum_{g=1}^{M} w_g \mathcal{N}(\mu_g, \mathbf{\Sigma}_g) \quad (4)$$

where $w_g$ are the mixture weights, $M$ is the total number of mixtures, $\mu_g$ are the mean vectors and $\mathbf{\Sigma}_g$ are the full covariance matrices. Here, $\mu_g$ and $w_g$ are the same as in (2) and (3).

#### *2.2.2. Noise subspace selection*

The AFA parameter $q$ in (1) defines the number of principal axes to retain, assuming the lower $(d-q)$ directions spans the noise-only subspace [7]. The maximum likelihood estimate of the noise variance $\sigma_g^2$ for the $g$-th mixture is given by,

$$\sigma_g^2 = \frac{1}{d-q}\sum_{i=q+1}^{d} \lambda_{g,i} \quad (5)$$

where $\lambda_{g,q+1} \cdots \lambda_{g,d}$ are the $d-q$ smallest eigenvalues of $\mathbf{\Sigma}_g$. We note that $q$ can be different for each mixture, and thus in the later sections, we denote it by $q(g)$.

#### *2.2.3. Compute the factor loading matrix*

The maximum likelihood estimation of the factor loading matrix $\mathbf{W}_g$ of the $g$-th mixture of the AFA model in (2) is given by [10],

$$\mathbf{W}_g = \mathbf{U}_{\mathbf{q}_g}(\mathbf{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I})^{1/2}\mathbf{R}_g \quad (6)$$

where $\mathbf{U}_{\mathbf{q}_g}$ is a $d \times q$ matrix whose columns are the $q$ leading eigenvectors of $\mathbf{\Sigma}_g$, $\mathbf{\Lambda}_{\mathbf{q}_g}$ is a diagonal matrix containing the corresponding $q$ eigenvalues, and $\mathbf{R}_g$ is a $q \times q$ arbitrary orthogonal rotation matrix. In this work, we set $\mathbf{R}_g = \mathbf{I}$.

#### *2.2.4. The AFA transformation*

The posterior mean of the *acoustic factors* $\mathbf{y}_n$ can be used as the transformed and dimensionality reduced version of $\mathbf{x}_n$ for the $g$-th component of the AFA model. This can be shown to be [10]:

$$E\{\mathbf{y}_n|\mathbf{x}_n, g\} = \langle \mathbf{y}_n|\mathbf{x}_n, g\rangle = \mathbf{A}_g^T(\mathbf{x}_n - \mu_i) \triangleq \mathbf{z}_{n,g} \quad (7)$$

where

$$\mathbf{A}_g = \mathbf{W}_g\mathbf{M}_g^{-T} \text{and} \quad (8)$$

$$\mathbf{M}_g = \sigma_g^2\mathbf{I} + \mathbf{W}_g^T\mathbf{W}_g. \quad (9)$$

The matrix $\mathbf{A}_g$ is termed the $g$-th *AFA transform*. Here, the original feature vectors $\mathbf{x}_n$ are replaced by the mixture dependent transformed vectors $\mathbf{z}_{n,g}$. It can be easily shown that $\mathbf{z}_{n,g}$ is normally distributed with a zero mean and diagonal covariance matrix $\mathbf{\Sigma}_{\mathbf{z}_g} = \mathbf{I} - \sigma_g^2\mathbf{\Lambda}_{\mathbf{q}_g}^{-1}$, demonstrating that $\mathbf{A}_g$ performs mean normalization and de-correlation. Conventionally, a feature vector $\mathbf{x}_n$ is aligned with a mixture $g$ that yields the highest posterior probability $p(g|\mathbf{x}_n, \Lambda_0)$, and the corresponding transformation is applied for dimensionality reduction [10]. However, as we observe the distribution of $\max_g p(g|\mathbf{x}_n, \Lambda_0)$ for our development data in Fig. 1(c), features are aligned with multiple Gaussians in most cases providing values of $\max_g p(g|\mathbf{x}_n, \Lambda_0) \sim 0.5$. Thus, we propose to apply the AFA transform using a probabilistic alignment and transform the UBM instead of retraining it. The new UBM is given by,

$$p(\mathbf{z}|\hat{\Lambda}_0) = \sum_{i=1}^{M} w_g \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{z}_g}) \quad (10)$$

### 2.3. Feature enhancement in AFA

Expansion of the transformation matrix $\mathbf{A}_g$ in (8) unveils the built-in enhancement operation it performs. Substituting the expressions of $\mathbf{W}_g$ and $\mathbf{M}_g$ from (6) and (9) in (8), we have:

$$\mathbf{A}_g^T = \mathbf{\Lambda}_{\mathbf{q}_g}^{-1}(\mathbf{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I})^{T/2}\mathbf{U}_{\mathbf{q}_g}^T = \mathbf{\Lambda}_{\mathbf{q}_g}^{-\frac{1}{2}}\mathbf{G}_g\mathbf{U}_{\mathbf{q}_g}^T \quad (11)$$

where we utilized the fact that, $\mathbf{U}_{\mathbf{q}_g}^T\mathbf{U}_{\mathbf{q}_g} = \mathbf{I}$ and introduced a diagonal gain matrix given by:

$$\mathbf{G}_g = \mathbf{\Lambda}_{\mathbf{q}_g}^{-\frac{1}{2}}(\mathbf{\Lambda}_{\mathbf{q}_g} - \sigma_g^2\mathbf{I})^{T/2}. \quad (12)$$

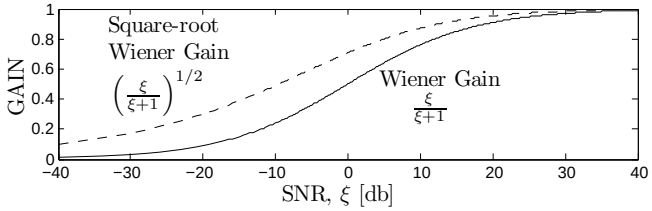



Figure 2: Input SNR [dB] ($\xi$) vs. Wiener gains. Wiener and square-root Wiener gains are shown with a solid (-) and dashed (- -) line, respectively.

Keeping aside the term $\mathbf{\Lambda_q}_g^{-\frac{1}{2}}$ in (11), we observe that the operation in (7) first computes the inner product of the mean normalized acoustic feature with the $q$ principal eigenvectors of $\mathbf{\Sigma_g}$, then for each $i$-th eigenvector direction, applies the gain function specified by the $i$-th diagonal of $\mathbf{G}_g$ in (12), which is actually a square-root Weiner gain function [11]. Defining the classic speech enhancement terminology *a priori* SNR $\xi$ as [7] $\xi = (\lambda_{g,i} - \sigma_g^2)/\sigma_g^2$, and using this to express the gain equations, we plot the gain functions against $\xi$ in Fig. 2. Thus, in addition to de-correlation and dimensionality reduction, the transform $\mathbf{A}_g$ also performs feature enhancement, assuming a noise variance $\sigma_g^2$. It maybe noted that conventional factor analysis techniques in super-vector space are also known to be representable using similar Wiener like gain functions as discussed in [12].

### 2.4. Feature variance normalization in AFA

The term $\mathbf{\Lambda_q}_g^{-\frac{1}{2}}$ in (11) normalizes the variance of the acoustic feature stream in the $i$-th eigen-direction, since $\lambda_{g,i}$ is the expected feature variance along this direction [13]. The AFA transformation performs this normalization in each mixture in addition to the enhancement mentioned in the previous section. This process is interestingly similar to the cepstral variance normalization frequently performed in the front-end except for its domain of operation.

### 2.5. Adaptive AFA dimension selection

Since the distribution of eigenvalues are different in each mixture, a unique value of $q(g)$ should be suitable in each case. This is illustrated in Fig. 1(b), where eigenvalues of three different mixture covariance matrices are shown. Typically we see an energy compaction of $\sim 70\%$, that is, the first $70\%$ eigenvalues account for $95\%$ of the total energy. But in other cases, the energy compaction can be very low or high, demanding that the AFA retained dimension to be low or high, respectively. Motivated by this observation, we develop a simple method of selecting the AFA dimension, $q(g)$. We first set the energy compaction ratio $E$ close to $0.9 \sim 0.97$, then compute the sorted eigenvalues $\lambda_{i,g}$ of $\mathbf{\Sigma}_g$ for each mixture $g$. The optimal dimension retaining $E\%$ energy is then calculated as:

$$q_E(g) = \min_q s.t. \frac{\sum_{i=1}^{q} \lambda_{g,i}}{\sum_{i=1}^{d} \lambda_{g,i}} > E. \qquad (13)$$

This method is denoted by "AFA-Var-En". As an alternate method, we use the effective rank estimation algorithm in [14], generally used for noise estimation in matrices, to select the AFA dimension. A threshold $\delta \in [0, 1]$ is set and the AFA dimension is obtained by:

$$q_\delta(g) = \min_q s.t. \left(\frac{\sum_{i=1}^{q} \lambda_{g,i}^2}{\sum_{i=1}^{d} \lambda_{g,i}^2}\right)^{1/2} > \delta. \qquad (14)$$

We denote this method as "AFA-Var-Rk". When the AFA dimension fixed for all mixtures, we denote the system by "AFA-Fix".

## 3. AFA integrated i-Vector system

In this section, we describe how the proposed method could be incorporated into the current state-of-the-art i-Vector system framework [4]. First, a full covariance UBM model, $\Lambda_0$ given by (4), is trained on the development data vectors. Next, the AFA dimension $q(g)$ for each mixture $g$ is set, and the noise variance $\sigma_g^2$ is computed using (5). The factor loading matrix $\mathbf{W}_g$ and transformation matrix $\mathbf{A}_g$ are then computed using (6) and (8), respectively. For each development utterance $s$, the zero order statistics is given by,

$$N_s(g) = \sum_{n \in s} \gamma_g(n) \text{ where } \gamma_g(n) = p(g|\mathbf{x}_n, \Lambda_0) \qquad (15)$$

following the standard procedure [2,4]. However, for AFA, the first order statistics $\hat{\mathbf{F}}_s(g)$ is extracted using the transformed features in the corresponding mixtures instead of the original features.

$$\hat{\mathbf{F}}_s(g) = \sum_{n \in s} \gamma_g(n) \mathbf{z}_{n,g} = \mathbf{A}_g^T \sum_{n \in s} \gamma_g(n)(\mathbf{x}_n - \mu_g) \qquad (16)$$

For estimating the total variability (TV) matrix, the standard procedure is followed [4] using the new UBM $\hat{\Lambda}_0$ given in (10), and statistics $\hat{\mathbf{F}}_s(g)$ and $N_s(g)$. It should be noted that, in this case the super-vector dimension reduces to $K = \sum_{g=1}^{M} q(g)$ from $Md$, and TV matrix size becomes $K \times R$. We define super-vector compression ratio $\alpha = K/Md$, measuring overall AFA compaction.

## 4. Experiments and Results

We perform our experiments on the male trials of NIST SRE 2010 telephone train/test condition (condition 5, normal vocal effort).

### 4.1. System Description

For voice activity detection, a phoneme recognizer [15] combined with an energy based scheme is used. 60-dimension feature vectors (19 MFCC +Energy + $\Delta$ + $\Delta\Delta$) are extracted, using a 25 ms window with 10 ms shift and Gaussianized using a 3-s sliding window. Gender dependent full and diagonal covariance UBMs with 1024 mixtures are trained on utterances selected from Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrollment data. For the TV matrix training, the same dataset is utilized. 400 dimensional i-Vectors are extracted, whitened and then length normalized [5]. For session variability compensation and scoring we use a Gaussian Probabilistic Linear Discriminant Analysis (PLDA) scheme with a full-covariance noise model [5].

### 4.2. Results using AFA transformation

We compare several AFA systems with our diagonal and full-covariance UBM based i-vector systems, denoted by "baseline diag-cov" and "baseline full-cov", respectively. The following AFA systems are built: "AFA-Fix" with $q(g) = 42, 48$ and $54$, "AFA-Var-En" with $E = 0.90, 0.95$ and $0.97$, and "AFA-Var-Rk" with $\delta = 0.99, 0.995$ and $0.997$. In this experiment, we fix the PLDA eigenvoice size $N_{EV}$ to 200. The results are shown in Table 1.

From the results, we observe that the AFA-Var-En system in general outperforms the basic AFA-Fix, even with similar compression ratio $\alpha$. The best results are obtained for AFA-Var-En ($E = 0.97$) with an EER of $2.0276\%$ obtaining a $15.37\%$ relative improvement from "Baseline full-cov". The AFA-Var-Rk ($\delta = 0.997$) method provided the best $\text{DCF}_{\text{old}}$ of $0.3786$. These results prove that the proposed AFA transformation is successfully able to reduce some nuisance directions in the feature space producing i-Vectors with better speaker discriminating ability. We also observe that AFA i-Vector systems are computationally less expensive compared to the baseline system, roughly by a factor of $\alpha$.

Table 1: Comparison between baseline i-Vector and AFA systems with respect to %EER, $DCF_{old}$ and $DCF_{new}$ for $N_{EV} = 200$. Percent relative improvement (%r) and super-vector compression ratio ($\alpha$) are also shown.

| System | | $\alpha$ | %EER/%r | $DCF_{old}$/%r | $DCF_{new}$/%r |
|---|---|---|---|---|---|
| Baseline full-cov | | 1.00 | 2.3959 | 0.1273 | 0.4534 |
| AFA-Fix | $q = 42$ | 0.70 | 2.36/1.42 | 0.13/1.57 | 0.44/3.46 |
| | $q = 48$ | 0.80 | 2.12/11.51 | 0.12/4.55 | 0.45/1.21 |
| | $q = 54$ | 0.90 | 2.25/5.96 | 0.12/4.47 | 0.44/3.90 |
| AFA-Var-En | $E = 0.90$ | 0.59 | 2.46/-2.56 | 0.13/-0.47 | 0.43/6.17 |
| | $E = 0.95$ | 0.72 | 2.13/11.12 | **0.11/12.96** | 0.40/11.55 |
| | $E = 0.97$ | 0.80 | **2.03/15.37** | 0.12/9.42 | 0.40/11.18 |
| AFA-Var-Rk | $\delta = 0.99$ | 0.53 | 2.40/-0.32 | 0.13/1.17 | 0.42/8.35 |
| | $\delta = 0.995$ | 0.62 | 2.25/6.19 | 0.12/5.26 | 0.42/7.39 |
| | $\delta = 0.997$ | 0.69 | 2.13/10.95 | 0.11/12.56 | **0.38/16.49** |

Table 2: Linear score fusion of baseline and AFA systems

| *Individual system performances* | | | | | |
|---|---|---|---|---|---|
| | **System** | $N_{EV}$ | **%EER** | **$DCF_{old}$** | **$DCF_{new}$** |
| (i) | Baseline full-cov | 200 | 2.3959 | 0.1273 | 0.4534 |
| (ii) | Baseline diag-cov | 200 | 2.4422 | 0.1243 | 0.4609 |
| (iii) | AFA-Var-En (0.97) | 200 | 2.0276 | 0.1153 | 0.4027 |
| (iv) | AFA-Fix (48) | 150 | 2.0591 | 0.1205 | 0.4600 |
| *Fusion system performances* | | | | | |
| 1 | Fusion of (i) & (iii) | | 1.9759 | 0.1080 | 0.3882 |
| 2 | Fusion of (i) & (iv) | | 2.0070 | 0.1103 | 0.4083 |
| 3 | Fusion of (i) - (iii) | | **1.8077** | **0.0993** | **0.3733** |

This is expected since the computational complexity of an i-Vector system is proportional to the super-vector size [16].

#### 4.2.1. *Fusion of multiple systems*

We pick four of our systems for fusion: (i) Baseline full-cov, (ii) Baseline diag-cov, (iii) AFA-Var-En ($E = 0.97$), and (iv) AFA-Fix ($q = 48$). The PLDA $N_{EV}$ parameter was set to 200 for systems (i)-(iii) and 150 for system (iv). Simple linear fusion was used with mean and variance normalization of scores to $(0, 1)$ for calibration. From the results presented in Table 2, fusion performance of (i) and (iii) clearly reveals that AFA and "baseline full-cov" systems have complementary information, as the EER and DCF values improve. The best result is achieved by fusing systems (i), (ii) and (iii), to obtain: EER = 1.807%, $DCF_{old}$ = 0.0993 and $DCF_{new}$ = 0.3733. Performance comparison of the systems (i), (ii) and the fusion is shown in Fig. 3 using Detection Error Trade-off (DET) curves. Here, again we observe the superiority of the proposed AFA-Var-En system compared to the baseline especially in the low false alarm region, whereas the fusion system shows better performance in the full DET curve range.

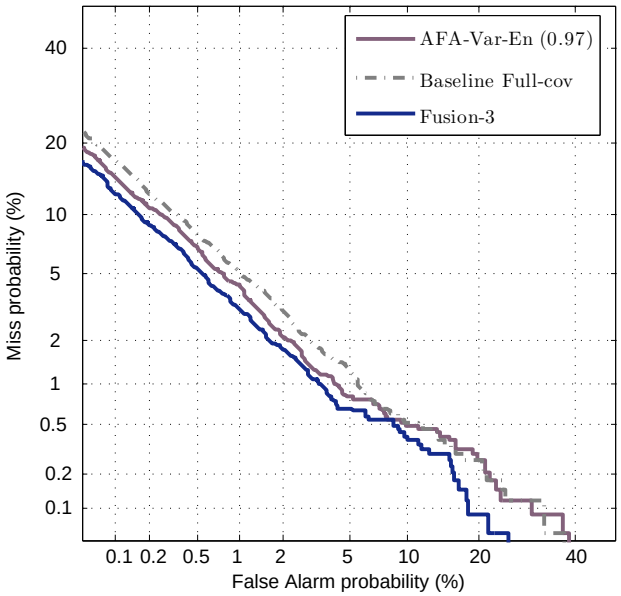



Figure 3: DET curves showing individual subsystems: AFA-Var-En ($E = 0.97$), Baseline full-cov, and Fusion of (i)-(iii) shown in Table 2.

## 5. Conclusions

In this study, we have proposed a factor analysis model for acoustic features to compensate for transmission channel mismatch in speaker recognition. Using the model, we have developed a mixture dependent feature transform that performs dimensionality reduction, de-correlation, variance normalization and enhancement, at once. Instead of a separate front-end processing, the proposed transform has been integrated within an i-Vector speaker recognition framework using a probabilistic feature alignment technique. Experimental results have demonstrated the superiority of the proposed scheme compared to the baseline i-Vector system.

## 6. References


[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19 – 41, 2000.

[2] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus Eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 99, pp. 788 – 798, May 2010.

[5] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-Vector length normalization in speaker recognition systems," in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 249 – 252.

[6] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. ICASSP*, Florence, Italy, Oct. 2011, pp. 4828 – 4831.

[7] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul 1995.

[8] B. Zhou and J. H. L. Hansen, "Rapid discriminative acoustic model based on Eigenspace mapping for fast speaker adaptation," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 4, pp. 554 – 564, July 2005.

[9] T. Hasan and J. H. L. Hansen, "Factor analysis of acoustic features using a mixture of probabilistic principal component analyzers for robust speaker verification," in *Proc. Odyssey*, Singapore, June 2012.

[10] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[11] S. Vaseghi, *Advanced signal processing and digital noise reduction*. Wiley, 1996.

[12] A. McCree, D. Sturim, and D. Reynolds, "A new perspective on GMM subspace compensation based on PPCA and wiener filtering," in *Proc. Interspeech*, Florence, Italy, Oct. 2011, pp. 145 – 148.

[13] T. Anderson, "Asymptotic theory for principal component analysis," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 122–148, 1963.

[14] J. Cadzow, "SVD representation of unitarily invariant matrices," *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. 32, no. 3, pp. 512 – 516, Jun 1984.

[15] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP*, vol. 1, May 2006.

[16] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proc. ICASSP*, Florence, Italy, Oct. 2011, pp. 4516 – 4519.